Advances in Intelligent Systems Research, volume 172
Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019)

Handling Numerical Features on Dataset Using Gauss Density Formula and Data Discretization Toward Naïve Bayes Algorithm

Mochammad YUSA[1*], Ernawati ERNAWATI[2], Yudi SETIAWAN[3], and Desi ADRESWARI[4]

[1, 2, 4] *Department of Informatics, Faculty of Engineering, University of Bengkulu,*
[3] *Department of Information System, Faculty of Engineering, University of Bengkulu,*
[1,2,3,4] *Jln. WR Supratman Kandang Limun Muara Bangkahulu Bengkulu Indonesia*
**Corresponding author: mochammad.yusa@unib.ac.id*

Abstract
Naïve Bayes is one of best classifiers in data mining. Naïve Bayes Algorithm either is used in some research areas. Besides having good performances, the algorithm can also handle numerical and categorical data values. This paper presents two ways of treating numerical features as a pre-process before implementing Naïve Bayes algorithm in classifying a dataset. First way is by implementing Gauss Density Formula. In second way, we treat the numerical features to be categorized manually by involving the experts. This study start from collecting data which contains numerical attributes in majority. Then dataset will be treated by using first way and second way. We validate the performance of algorithm by using 10-Fold Cross Validation. The considered performances in this research are accuracy, precision, and recall. The result shows that treating numerical features using Gauss Density Techniques outperforms the treatment by discretizing numerical features of nominal values. First way obtains 80% accuracy, 80,61% of precision average, and 80,41% of recall average value while the second way reaches 65% of accuracy, 63,95% of precision average, and 66,43% of recall average.

Keywords: *Gauss Density, data discretization, Naive Bayes, data mining, classifiers*

Introduction

Naïve Bayes Algorithm forms a powerful classifiers since it was implemented in some research areas by some researchers. Reference [1] used the algorithm to classify texts. In art field, reference [2] applies Naive Bayes Algorithm to classify song emotion from the lyrics. In another area, some researchers use the naive Bayes classifiers to help some scientists identify the kinds of disease that are suffered by patients [3]. Jyothi and Bhargavi [4] apply Naïve Bayes Algorithm to predict agricultural land soils. Since it was well-known, the algorithm was used in some areas such as document classification, art, health, and agriculture.

Besides becoming the famous classification algorithm, Naïve Bayes also obtains better performance than any classification algorithms. Reference [5] compares some classification algorithms towards dataset of readmission diabetic patients and find that Naïve Bayes outperforms K-NN and C4.5 Decision Tree. Another research proved that Naïve Bayes classifier has better results than the C4.5 classifier [6]. Amra and Maghari also implement Naïve Bayes and K-NN as methods to predict student performance and find out that Naïve Bayes still achieves better accuracy than the K-NN method [7] either Reference [8] finds out that Naïve Bayes has a strong performance.

Knowledge discovery of data mining consists of many primary processes. After Data Understanding process, the next stage is data preparation. Preparation data is one of important steps in searching knowledge from a raw of data. Some of them are feature selection, missing data analysis, data discretization, data transformation, etc. Previous studies about implementing Naïve Bayes algorithm commonly apply standard deviation to treat numerical values on dataset [4, 6, 7] while Naïve Bayes may have good performances if the data value on numerical features is discretized into nominal values [9]. Until now, we have not found yet the research that practiced the treatment for handling numerical values as preprocessing stage before implementing Naïve Bayes algorithm. In other hand, some researchers say that preprocessing process is crucial in efforts of improving algorithms' performances [10,11,12]. The aim of this study is to investigate the effect of Naïve Bayes performances based on numerical features treatment.

Methods

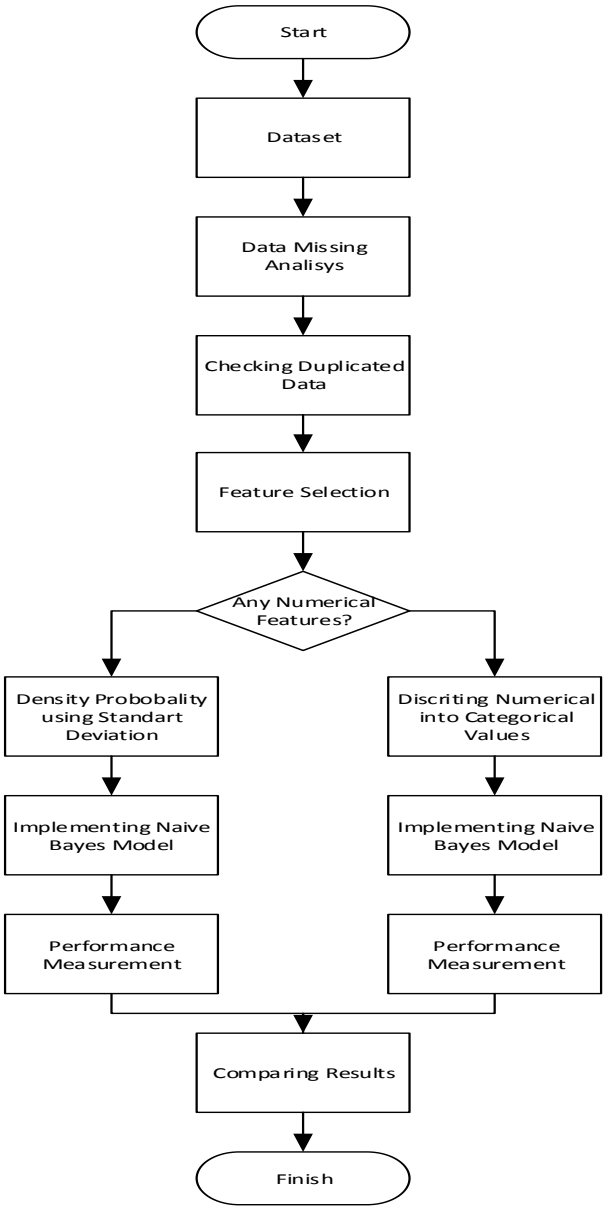

Figure 1. Method of the research

Fig. 1 represents the research method. This research starts from preparing data consisting of numerical attribute types in majority. The dataset was taken in The Indonesian Red Cross (IRC) in Bengkulu got dataset containing 100 records and nine attributes. Of the nine attributes, there are three attributes that are nominal and the rest are numerical. So, this dataset is very suitable to be used to calculate Naïve Bayes performance in handling attributes that consist of numerical values in majority.

The first stage of this research is data preparation. At this stage, there are three sub-steps that we are going to do, such as analyzing missing data, checking duplicated data, and feature selection. They will be used for the pre-process of finding knowledge based on the Naive Bayes method. After the step is completed, the next step is to implement the technique with a different treatment of the numerical attributes. The first step is to use calculations using the Gauss Density formula while the second is by discretizing the numerical attributes into nominal values. After completion, dataset will be implemented in overcoming Naïve Bayes performance. The parameter used is the accuracy of the validation process using the k-fold cross-validation technique with a subset value of ten. After the results are obtained, the results will be represented based on the accuracy performance based on the treatment of attributes that have numerical values.

Naive Bayes Algorithm

Naïve Bayes Classification is a statistical classifier which can be used to predict posterior probability of class. According to Wu and Kumar, Naïve Bayes is one of top ten classifiers on data mining methods. The method that is used on this technique utilizes a branch of math which has widely been known by probability theory to search maximum likelihood on class or label by scanning the frequencies of each class on data training [13]. Naïve Bayes Theorem can be formulated on (1):

$$P(H|X) = \frac{P(X|H)}{P(X)}.P(H) \qquad (1)$$

Where X is data with class that will be classified, H represents the defined data hypothesis. P (H|X) is a prior probabilities of H based on X condition (posterior probabilities). P (X|H) indicates the likelihood of X referred to condition on hypothesis of H and P(X) represents the probability of X. Naïve not only it can handle categorical values on attributes, but Naïve Bayes also can handle continuing values on attributes. Here is the formula to calculate posterior probabilities in effort to handle numerical attribute by using standard deviation (2) that is extended from Gauss Density Formula.

$$P(X|Y) = \frac{1}{\sqrt{2\mu\sigma}}.exp^{\frac{-(x-\mu)^2}{2\sigma^2}} \qquad (2)$$

P (X|Y) indicates a probability X attributes to Y target class. Otherwise μ is mean value from summarization of numerical attributes and σ forms the principal deviation standard stating variants for all attributes.

Dataset

The dataset used in this study is a dataset from IRC in Bengkulu section. We involve experts to consider influential data in classifying donor eligibility. This dataset consists of nine attributes / features which include eight bound attributes and one attribute label or target class. The amount of records used in this study is 100 records which have been validated by a member of the IRC. These attributes can be seen in Table 1.

Table 1. The attribute list of the Dataset

Num.	Feature name	Data Type	Value
1	Age	Numerical	19, 20, 67,... in years-old
2	Weight	Numerical	51, 49, 67, ... In kilogram
3	Hemoglobin	Numerical	12.05, 15.00, 15.01, ... In gram
4	Gender	Binominal	Male or Female
5	Systolic Blood Pressure	Numerical	100, 110, 120, 130, ... in mmHg
6	Dystolic Blood Pressure	Numerical	60, 70, 80, 90, ... in mmHg
7	pulse rate per minute	Polynomial	0-49, 50-100, >100
8	Last Three-month donor history	Binominal	"Yes" value is for volunteers who have donated in last three months and "No" is for volunteers who have not
9	Donor Eligibility	Binominal	Yes or No contains data values of label

DATA PREPARATION

After collecting data, next step is analyzing the attributes on dataset as been shown in Fig. 1. We discussed with the members of IRC related to the blood donor features which classification will be used to implement the Naïve Bayes Algorithm. Selecting the right attributes before implementing Algorithm Model is very important. The method that we used in this study is by involving the expert related to blood donor activity to select the features that are surely influential to target class. According to the experts, they said that human pulse is useless because it tends to be fluctuated and does not affect to the activity of selecting candidate of blood donor. Then, other features which are not chosen are "Last Three-month donor history". It is because it will result the constant result. So there are two features which are not selected in this study. However, as the result of feature selection, "human pulse" and "Last Three-month donor history" attribute are not considered to the final dataset. Table 2 shows a list of comprehensive features that are used after the feature selection stage is done.

Table 2. Final Dataset

Num.	Feature name	Data Type	Value
1	Age	Numerical	19,20,67,... in years-old
2	Weight	Numerical	51,49,67,.... In kilogram
3	Hemoglobin	Numerical	12.05, 15.00, 15.01,....

			In gram
4	Gender	Binominal	Male or Female
5	Systolic Blood Pressure	Numerical	100, 110, 120, 130,... in mmHg
6	Dystolic Blood Pressure	Numerical	60, 70, 80, 90,... in mmHg
7	Donor Eligibility	Binominal	Yes or No contains data values of label

While feature selection is done, the next step is the missing value analysis. The results of checking missing data indicate that there are no blank data. The final step after checking the missing data also shows that no duplicate data rows so that all data in the dataset are unique. Finally it can be concluded temporarily that there are seven features / attributes and 100 rows of data that will be the final dataset for next step of this study.

EXPERIMENTAL RESULTS

In this experiment, we use the Rapid Miner Machine Learning Tool version 8.1. In the first experiment, we will use the Gauss Density formula to find mean and standard deviation score as effort to handle numerical features. At the second experiment, we involve the expert to manually categorize the numerical features into nominal features.

After that, we implement Naïve Bayes algorithm to predict the posterior probabilities.

USING STANDARD DEVIATION

First experiment starts from importing data that has been processed at the data preparation stage. Then the set role function is added to determine the features that are labeled as the target class. In this research, the attribute that becomes the label parameter of the model is the donor classification attribute. Validation model which we used in this study is k-fold cross-validation with a fold value of ten.

The results of the calculation of the mean and standard deviation are then used to calculate all possible probabilities based on Eq. (1). The results of the Naive Bayes model calculation are represented in a Confusion Matrix as shown in Table 3 bellow.

Table 3. Matrix Confusion using Gauss Density Formula

	True. No	**True. Yes**
Pred. No	41	14
Pred. Yes	6	39

We can calculate the accuracy of Naïve Bayes Algorithm from confusion matrix that shows in Table 3. Result shows that Naive Bayes has 80% of accuracy if it is treated by using Gauss Density Formula. The accuracy performance indicates that it is categorized in good classification. Another performances such as Precision and Recall can be seen in Fig. 2.

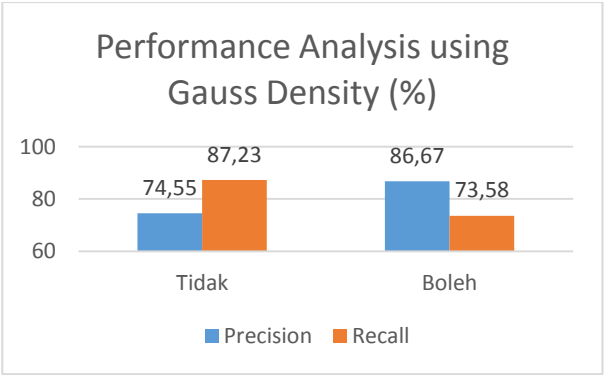

Figure *2*. Recall and Precision in percent

No

Fig. 2 indicates the results based on precision and recall performance of the algorithm. The Precision performance parameters obtained from the results of this experiment are quite good. Precision score to the prediction of "No" is classified as good with value of 74.55% and class precision for the prediction of "Yes" is 86.67%. The average class precision of the two labels is 80.61%. Then for the label "No", the recall value is 87.23% while for the label "Yes" is 73.58%. The average of recall performance of the two labels is 80.41%. It indicates that the recall level is classified as very well performance. Based on the results of experiment, in can be said that the Naive Bayes algorithm has a very good performance in handling numerical data value using Gauss Density Formula.

By Discretizing Numerical attributes to categorical

In the second experiment, we involved the experts from IRC to help this study in grouping the numerical data into nominal data. After giving the dataset, they give us the list of value (Table 4). Then, we transform the numerical data into nominal data.

Table 4. List of feature value

No	Attribute	Nominal
1	Age	Teenager, adult, middle-aged, elderly
2	Weight	{45-54}, {55-74}, and {>75}
3	Hemoglobin	{>15}, {<12,5}, and {12,5-15}
4	Systolic Blood Pressure	{<100}, {100-130}, and {>130}
5	Dystolic Blood Pressure	{55-80} and {>80}

After the numerical data has been clustered, the next process is importing the data Rapid Machine Learning. And so, we apply the Naïve Bayes algorithm and validates 10-fold Cross Validation to determine the Matrix Confusion. Table 5 indicates the result represented by the confusion matrix.

Table *5*. Confusion Matrix using Data Discretization

	True. No	True. Yes
Pred. No	21	9
Pred. Yes	26	44

Based on the results which is shown in Fig. 3, accuracy performance of the algorithm reaches 65%. There is 15% gap if it is compared to previous experimental result. The recall performance for the prediction "Yes" is 62.86% and "No" is 70%. And so, the average of recall performance is 66.43%. Third performance result also shows that Precision Value for the label "Yes" is 83.03% while the label "No" is 44.86%. The average precision value is 63.95%. This shows that the level of precision, recall, and accuracy were decreasing had it compared to the treatment of numerical attributes using the standard deviation formula.

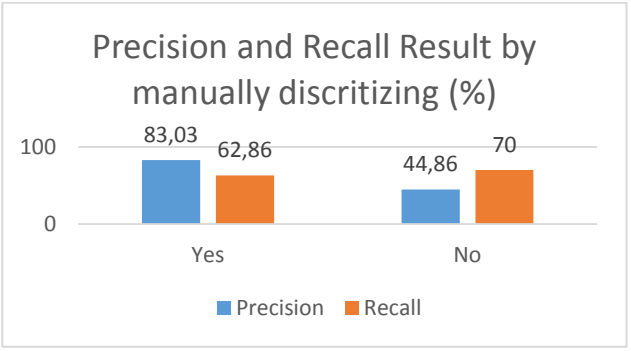

Figure 3. Precision and Recall Result by Discretizing Manually

Naive Bayes performance containing accuracy, recall, and precision using different treatment of numeric attributes in the dataset is shown in Table 6. Based on the results of the two experiments, it can be seen that applying Gauss Density technique has better accuracy, mean precision, and recall than using manual data discretization in handling numerical attributes.

Table 6. Results comparison

Numb	Treatment	Performance (%)		
		Accuracy	Precision Avg.	Recall Avg.
1	Using Gauss Density	80	80,61	80,41
2	Discretizing Manually	65	63,95	66,43

CONCLUSION

The main objective of this study is to investigate performances of the Naive Bayes algorithm by treating numerical attributes in different ways. The first experiment is by using standard deviation and mean as the basic computational to predict the posterior probability. The second experiment is by discretizing numerical attributes into nominal value attributes manually. The results shows that treatment using standard deviations have better performance from the way using data discretization method referring to accuracy, recall, and precision value. The further development for this study is to apply data discretization functions using statistical calculations such as based on technique of binning, sizing, entropy, and others.

ACKNOWLEDGMENT

Thanks to the Rector, Dean of Engineering Faculty, Department Head of Informatics of University of Bengkulu who have been financially supported this research.

REFERENCES

[1] S. Xu, „Bayesian Naive Bayes classifiers to text classification,“ *Journal of Information Science*, Vol. 44, No. 1, p. 48–59, 2018.

[2] Y. An, S. Sun und S. Wang, „Naive Bayes Classifiers for Music Emotion Classification Based on Lyrics,“ in *ICIS 2017*, Wuhan, 2017.

[3] S. A. Pattekari und A. Parveen, „Prediction System For Heart Disease Using Naive Bayes,“ *International Journal of Advanced Computer and Mathematical Sciences*, Bd. 3, Nr. 3, pp. 290-294, 2012.

[4] S. Jyothi und P. Bhargavi, „Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils,“ *IJCSNS International Journal of Computer Science and Network Security* , Vol. 9, No. 8, 2009.

[5] M. Yusa und E. Utami, „Classifiers evaluation: Comparison of performance classifiers based on tuples amount,“ in *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Yogyakarta, 2017.

[6] E. N. Azizah, U. Pujianto, E. Nugraha und Darusalam, „Comparative performance between C4.5 and Naive Bayes classifiers in predicting student academic performance in a Virtual Learning Environment,“ in *4th International Conference on Education and Technology (ICET)*, Malang, 2018.

[7] I. A. A. Amra und A. Y. A. Maghari, „Students Performance Prediction Using KNN and Naïve Bayesian,“ in *8th International Conference on Information Technology (ICIT)*, Amman, 2017.

[8] S. Hassan, M. Rafi und M. S. Shaikh, „Comparing SVM and NaIve Bayes Classifiers for Text Categorization with Wikitology as knowledge enrichment,“ in *IEEE 14th International Multitopic Conference*, Karachi, 2011.

[9] K. Lavangnananda and S. Chattanachot, "Study of Discretization Methods in Classification," in *9th International Conference on Knowledge and Smart Technology (KST)*, Chonburi, 2017.

[10] J. Chen, H. Huang, S. Tian and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, p. 5432–5435, 2009.

[11] P. Chandrasekar und K. Qian, „The Impact of Data Preprocessing On the Performance of Naïve Bayes

Classifier,“ in *40th Annual Computer Software and Applications Conference*, Atlanta, 2016.

[12] C. Kim und K.-B. Hwang, „Naive Bayes Classifier Learning with Feature Selection for Spam Detection in Social Bookmarking,“ Ministry of Knowledge Economy of Korea, Seoul, 2008.

[13] X. Wu and V. Kumar, "The top ten algorithms in data mining," *International Statistical Review*, vol. 78, no. 1, p. 158–158, 2009.